ENGINEERING-PSYCHOLOGY RESEARCH LABORATORY

University of Illinois at Urbana-Champaign

Technical Report EPL-83-1/ONR-83-1

June, 1983

A Comparison of Verbal and Graphical Information Presentation in a Complex Information Integration Decision Task

Christopher D. Wickens

Brad D. Scott

Prepared for

Office of Naval Research

Engineering Psychology Program

Contract No. N-000-14-79-C-0658

Work Unit No. NR 196-158

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER EPL-83-1/ONR-83-1	2. GOVT ACCESSION NO. AD-A130 482	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Comparison of Verbal and Graphical Information Presentation in a Complex Information Integration Decision Task		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Christopher D. Wickens & Brad D. Scott		8. CONTRACT OR GRANT NUMBER(s) N000-14-79-C-0658
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Illinois at Urbana-Champaign Department of Psychology 603 E. Daniel St. Champaign, IL 61820		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 196-158
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Engineering Psychology Program 800 N. Quincy St. Arlington, VA 22217		12. REPORT DATE June 1983 13. NUMBER OF PAGES 41
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release. Distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Decision making, display, compatibility, information, reliability

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes an experiment conducted to evaluate the relative merits of verbal as opposed to spatial-graphical display formats in presenting sequential information to subjects in a tactical decision making task. The task required subjects to integrate a series of information messages bearing on the likelihood that one of two hypotheses pertaining to tactical maneuvers was in effect. Each information source could vary in its diagnosticity and its reliability. These variables contribute independently to the total valence

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

or information valence of the cue. Subjects integrated information in problems of either 6, 8, or 10 cues, presented at a slow or fast speed, in either a verbal (numerical) or spatial (graphical) format. After each problem, subjects made a choice of the most likely hypothesis accompanied by an analog judgment of their confidence in that choice.

The results were examined from two perspective: (1) From the perspective of human engineering guidelines, the data indicated that subjects' decisions were more accurate using the spatial display. This finding supports the principle of S-C compatibility stating that the analog operations on which the judgments were based would be best served by spatial displays. The spatial advantage was enhanced when the cues were delivered at a slower speed, imposing greater demands on working memory. (2) The data for both displays were analyzed from the perspective of different models of probablistic information integration. With both display configurations, subjects tended to apply an absolute, rather than a relative judgment of cue reliability. They did not appear to be influenced by either recency or primacy (anchoring), but appeared to down weight differences in the reliability of information sources, relative to the optimal.

A Comparison of Verbal and Graphical Information Presentation in a Complex Information Integration Decision Task

Christopher D. Wickens and Brad D. Scott

Abstract

This report describes an experiment conducted to evaluate the relative merits of verbal as opposed to spatial-graphical display formats in presenting sequential information to subjects in a simulated C^3 tactical decision making task. The task required subjects to integrate a series of information messages bearing on the likelihood that one of two hypotheses pertaining to tactical battlefield maneuvers was in effect. Each information source could vary in its diagnosticity and its reliability. These variables contribute independently to the total valence or information value of the cue. Subjects integrated information in problems of either 6, 8, or 10 cues, presented at a slow or fast speed, in either a verbal (numerical) or spatial (graphical) format. After each problem subjects made a choice of the most likely hypothesis accompanied by an analog judgment of their confidence in that choice.

The results were examined from two perspectives: (1) From the perspective of human engineering guidelines the data indicated that subjects' decisions were more accurate using the spatial display. This finding supports the principle of S-C compatibility stating that the analog operations on which the judgments were based would be best served by spatial displays. The spatial advantage was enhanced when the cues were delivered at a slower speed, imposing greater demands on working memory. (2) The data were analyzed from the perspective of different models of probablistic information integration. In two respects subjects tended to optimal behavior: they tended to apply an absolute, rather than a relative judgment of cue reliability, and they did not appear to be influenced by either recency or primacy (anchoring). However, they did appear to down weight differences in the reliability of information sources, relative to the optimal.

Introduction

Tactical decisions are made by military commanders operating within command, control, and communications (C^3) systems. C^3 is defined as a closed-loop man-machine system designed to facilitate decision making authority and direction by a properly designated personnel. Currently, C^3 systems have a "real time" characteristic where commanders operate under conditions of time stress and uncertainty. The amount of available information will vary greatly and typically has a short life cycle. Samet, Weltman, and Davis (1976) assert that computer-based military systems for C^3 operations have increased the rate and density of information flow to such an extent as to overwhelm the commander. They assert that more information may in fact degrade, rather than enhance the absolute level of decision making performance. Loo (1981) characterizes the current state of affairs as "crisis management." The decision making performance of the military commander is the essence of the C^3 system. Consideration of human performance limits in the iterative process of C^3 system design is critical to system performance. Several approaches are being taken to improve decision making performance within the C^3 system. Of these, probably the most successful involves the development of decision aids.

Decision Aids

Decision aids can be implemented at different levels within the C^3 system and have inherently different functions. For example, linear models can be used to augment or replace the decision maker (Dawes, 1979). Proper linear models can be used in a normative sense to integrate information for the decision maker. Paramorphic models are created by modeling the decision maker and are considered an improper linear model in the sense that the derived weights are non-optimal. Bootstrapping involves replacing the decision maker with a paramorphic model. It has been demonstrated in clinical studies that all three models can out-perform the decision maker using various criterion variables (Dawes & Corrigan, 1974; Dawes, 1979; Goldberg, 1970). Dawes (1979) addresses questions raised about the technical, psychological, and ethical problems associated with the implementation of these linear models. Several issues raised here remain particularly relevant to the application of linear models in tactical C^3 systems.

Samet, Weltman, and Davis (1976) take yet another approach and make a solid case for an adaptive computerized system to control information flow so as to best match overall system and human capabilities. They assert that the multi-attribute information utility model is superior to the decision maker at selecting the optimal amount and type of information in inferential decision making tasks. They propose that this approach will increase the efficiency and effectiveness of the decision maker.

The approach of the present study is somewhat different from those discussed above. We are concerned with the design of information displays that will enhance decision making performance in probablistic information integration tasks. The general task setting we consider is one in which the operator must integrate a series of sequentially presented information cues that vary in their diagnostic value with regard to a set of hypotheses. The operator uses this information, optimally to revise in working memory a continuous analog "scale" of confidence that the evidence favors one or the other hypothesis. Such a task imposes limitations of at least two general classes: On the one hand, several investigators have documented a wide variety of cognitive limitations related to such heuristics as anchoring, representativeness, and the non-optimal weighting of sequential cues (e.g., Kahneman & Tversky, 1973; Einhorn & Hogarth, 1981; Slovic, Fischhoff, & Lichtenstein, 1977; Lopes, 1982). On the other hand, some limitations may be perceptual in nature, describing bottlenecks between the way the information is physically formatted and the optimal analog model that is to be maintained in working memory. The following pages will discuss in turn research related to information integration and to stimulus processing.

Information Integration

An abundance of empirical evidence has demonstrated that human judgment often does not reflect an optimal normative model. Tversky and Kahneman (1973), Lyon and Slovic (1976) and others have shown that judges neglect base rate information. Tversky and Kahneman have demonstrated that cognitive heuristics such as "representativeness," "anchoring and adjustment," and "availability" result in systematic biases in judgment. Wickens (1983) has summarized a series of examples in which subjects tend to ignore differences in the reliability of information sources, when these are integrated in multi-element decision tasks. These phenomena have sometimes been faulted for being highly problem-specific and of insufficient magnitude to yield a priori empirical predictions (Bar Hillel, 1980; Wallsten, 1977, 1980). Nonetheless, there have been a sufficient number of studies demonstrating that systematic biases do exist in human judgment, resulting in deviations from normative models to suggest that these could represent an important source of difficulty in the C^3 environment.

Lopes (1982) has contrasted three models of information integration used by subjects in different experimental settings. Each describes the manner in which an internal subjective response, r, is attained and updated from the value of two or more cues or sources of information. These are the multiplying model, the averaging model, and the adding or relative weighting model.

Multiplying models, where $r_{ij} = s_{Ai}\, s_{Bj}$, have been widely studied. If an integration process adheres to this model, then the effect of one S variable is magnified by increasing levels of the other. In many of its applications to decision theory, the two stimulus or cue values represent a weight, usually associated with a probabilistic value, and a scale value associated with the content of the information. Thus changes in the value of an information source are assumed to have progressively more impact on R, as the weight of that source is increased.

Averaging models assume that the information from two or more cues will simply be averaged along some internal continuum, with each cue "pulling" the average toward its particular value. This latter characteristic is important. If a series of cues are being integrated over time, the averaging model says that each new cue will always adjust the running average toward the value of that cue, i.e., between the current average and the new cue value. While an averaging model is often appropriate in some circumstances (e.g., as the subject is talleying the average value of a series of numbers), Lopes emphasizes that, when integrating probablistic information the technique is clearly inappropriate. Consider for example a Bayesian inference task in which belief in one hypothesis is held to a relatively strong extent, and new evidence which only weakly supports the same hypothesis is delivered. The averaging model predicts that the new belief will be adjusted downward toward the neutral point, reflecting a weighted average of the old belief and the new evidence. Yet this is non-optimal. Any evidence, no matter how weak, in favor of that hypothesis should cause a shift toward greater belief. Lopes, however, reports that many people demonstrate this same non-optimal bias, applying averaging when a Bayesian inference is warranted.

The third class of models then, of which the Bayesian inference is a specific case, are the relative ratio models. These are optimal for Bayesian tasks in which there are two polar alternatives. Each sampled cue will move the current average in the direction according to the hypothesis supported, with a magnitude given by the diagnostic value of the stimulus. Only stimuli with no information value at all will fail to move the current pointer. Unlike the averaging model, the current pointer will only be moved toward a new piece of sampled evidence if the evidence is more extreme than the value of the current pointer. In essence then, new information is added to the integration of previous information, rather than being averaged with that information.

It is clear from a review of the literature on information integration that the model of prediction is largely dependent upon the problem situation and experiment 1 design. Lopes (1982) asserts that the difference in data that exhibit relative ratio procedure and data that

exhibit an averaging procedure can be explained in terms of the subjects' representation of the response scale and the order in which stimulus features are to be processed. In this respect, Lopes was able to make subjects more normative Bayesian information integrators by changing their adjustment strategies.

Each model has associated with it a series of weights, attached to each source of information that dictates how much each new item of information will impact on the running estimate. Therefore another characteristic that underlies models of information integration concerns the weights that are applied to different sources of evidence. Three characteristics of this weighting scheme are relevant to the present research: (1) How weights are assigned according to the order of cue arrival (the issue of primacy and recency or serial position), (2) How weights are assigned according to the nature of the information, (3) Whether absolute or relative weights are employed.

(1) Serial position effects: Lopes (1982) has demonstrated the manner in which the weights assigned to stimuli or information sources arriving in sequence dictate various serial position effects. In particular, heavy weightings assigned to the first arriving cues indicate anchoring or primacy: A relunctance to adjust current estimates in light of newly arriving information. Heavy weights assigned to the final cues suggest recency in the integration process. The potential role of these two biases will be examined in the present experiment.

(2) Differential weightings to different cues. Following the multiplicative model of information integration, the reliability of a piece of information should optimally be given the same weight as the diagnosticity of that information in choosing between hypotheses (i.e., the extent to which the value of the cue is more likely under one hypothesis than the other) (Johnson, Cavanagh, Spooner, & Samet, 1973). These two should multiply to derive the total information "worth." Yet there is evidence that when several sources of information differing in both diagnosticity and reliability must be integrated, subjects tend to discount differences in reliability of information sources, treating all sources as if they were fully reliable, and focussing attention instead more exclusively on diagnosticity (e.g., Kanarick, Huntington, & Peterson, 1969; Kahneman & Tversky, 1973; Schum, 1975). We shall be investigating this "as if" heuristic.

(3) Absolute versus relative weighting. The issue here concerns the extent to which people weigh the strength of evidence for or against a particular hypothesis relative to the total amount of evidence presented. Will the subject for example, who is confronted with a piece of evidence favoring a hypothesis by 70/30 odds be more inclined to believe that hypothesis true if this was the only evidence viewed, than if he had

previously viewed 10 pieces of evidence whose net effect was neutral (i.e., favored neither hypothesis). If the answer is affirmative, then the subject is employing some form of relative weighting scheme. The second case is seen by the subject to provide weaker evidence because the 70/30 odds were only obtained after 11 cues. In the first case, the same odds were attained with only a single case. On the other hand, if the subject views the two situations as providing equivalent evidence then he is following an absolute weighting scheme. Applying optimal Bayesian procedures, the posterior odds equal the prior odds multiplied by the likelihood ratio, and the question of how many cues were required to attain those prior odds is immaterial. This then is another issue we shall address.

The problem situation and experimental design of an investigation by Fleming (1970) is particularly relevant to our study. In his experiment subjects processed conflicting information in a tactical decision making task. Decision making performance was compared to a Baysian inference model whereby probabilities from successive information sources should, normatively, be multiplied to arrive at an overall probability for each of three alternative hypotheses. Fleming concludes that rather than obeying a normative model the majority of subjects used an adding model. Furthermore, all subjects that did not receive feedback were reported to have used an adding model. It is not clear what method of analysis Fleming used in concluding that his data exhibited an adding model of information integration.

A descriptive adding-multiplying model has been utilized to represent performance in the current study. The descriptive model of information integration is illustrated in Figure 1. The specific stimulus dimensions, reliability and diagnosticity, will be defined at a later point. Models A and B both present the same information, however, Model B requires only half the processing of Model A, i.e., only the differences between the information values S_{iA} supporting Hypothesis A and S_{iB} supporting Hypothesis B are given in Model B. Model B is more efficient and represents the actual form of the adding-multiplying model used in the current study. These models are similar to the adding model found in Fleming's (1970) study in that information integration across cues is an adding process. The optimum values of individual cues were given in Fleming's study while they are determined via a multiplicative process in our model. It is important to note that the model is not complete in that it does not describe a difference judgment. That is, at the end of the trial the subject has a trial value for each competing hypothesis, Hypothesis A and Hypothesis B, and must now differentiate between the two and decide which is the most likely hypothesis. This process will be discussed below in the experimental predictions.

Stimulus Structure

In the present experiment our subjects are required to integrate a series of cues each of which differed in reliability and diagnosticity. As noted above, these two values should optimally be multiplied to produce a new quantity, the total information value, worth, or valence of the cue, and it is these valences we ask subjects to derive and integrate across cues. In the "verbal" display condition, the reliability and diagnosticity values are presented numerically. In the spatial display, they are presented graphically as the height and base of a rectangle, respectively. The spatial display has two potential advantages over the verbal.

(1) According to the principle of stimulus/central-processing/response (S-C-R) compatibility, outlined by Wickens, Sandry, and Vidulich (1983), tasks that demand spatial/analog processes in working memory will be best served by visual spatial displays and more poorly served by verbal displays (either speech or print). Since the present task requires that the subject update a continuous scale of confidence in working memory with each added cue, the S-C-R compatibility theory predicts better performance with the graphical display.

(2) An added benefit of the format of the graphical display is that the height and width (diagnosticity and reliability) of each rectangle cue are combined in such a way as to produce a new dimension--area--that is directly equal to the cue valence measure subjects are supposed to integrate. A series of investigations indicate that these two dimensions are "integral" and so are combined "automatically" and holistically by the perceptual system to generate a direct perception of rectangular area (Garner, 1974; Garner & Felfoldy, 1970; Lockhead, 1979). Hence, with the spatial display, the subject does not need to engage in the conscious, cognitively loading multiplication process to combine the two dimensions and derive the valence measure.

There is however one potential drawback to the use of the holistic analog display that may lead to systematic biases. Smith (1969) has argued that the perception of rectangle size is influenced not only by area, but by perimeter as well. This conclusion accounts for Anderson and Weis's (1971) observation that the size of highly eccentric rectangles is consistently overestimated. Hence, elongated rectangles will be judged as larger than squares of the same area, because the perimeter of the former is greater. If this bias operates, then subjects will tend to overestimate the valence of cues in which reliability and diagnosticity are negatively correlated (producing elongation), relative to the square-producing cases in which the two variables covary in a positive fashion.

Experimental Predictions

Experiment 1

A series of decision problems were designed in the context of a tactical battlefield scenario. The subject was designated as a commander, responsible for defending an area through which an attack from a ficticious threat force was eminent. Threat force doctrine as well as terrain features in the area of operations dictated that the attack would come along a narrow front from either the north (H_1) or south (H_2) of their sector. It was the subject's duty to analyze the available intelligence and decide which avenue of approach, north or south, the threat force would take. Appendix A demonstrates this scenario. The information for each hypothesis in each problem was presented sequentially from several sources of information or cues. Each source conveyed information for one of the two possible hypotheses. The worth of each source was determined by two dimensions. Reliability (i.e., air reconnaissance report, reliability = .80) and diagnosticity (i.e., destruction of obstacles to south, D = .70). Subjects were instructed to evaluate the information presented and decide which hypothesis concerning future threat force actions was most likely to occur. The effects of five parameters on decision accuracy and confidence were studied within the framework of descriptive Model B presented in Figure 1. Predictions of these effects and how they bear on the model of information integration are now discussed.

1. Figures 2 and 3 illustrate the verbal and spatial code formats of the information display, respectively. It was predicted that the spatial code format which utilizes dimensional integrality would enhance decision accuracy. This prediction as noted above is based both upon the principle of S-C compatibility, and the integral "configural" nature of the rectangular object display.

While major emphasis of the present experiment was placed on the distinction between verbal and analog display formats, we were also interested in the influences of four additional decision problem variables, both in their own right, and as they might modify the effect of display format. These are described as follows.

2. The time available to process each cue in a decision problem was varied. The main effect of this variable was of less interest than were the interaction effects of this variable with cue coding. We were interested in how any advantages of the spatial display might be modulated by varying information rate. On the one hand, increased rate produces an increased degree of time-stress--presumably a detrimental effect. On the other hand, this may be balanced by the fact that the faster rate imposes less of a burden on working memory for the integration of successive cues. Because of these two counteracting trends, we were unable to predict a priori the ultimate effect of this variable.

Model A

Information Value

$$\begin{matrix} & \text{Hyp A} & \text{Hyp B} \\ \text{Reliability} & & \\ \left[r_1, r_2, r_3, \ldots, r_n\right] & \begin{bmatrix} S_{1A} & S_{1B} \\ S_{2A} & S_{2B} \\ S_{3A} & S_{3B} \\ \vdots & \vdots \\ S_{nA} & S_{nB} \end{bmatrix} & = \begin{bmatrix} \sum_{i=1}^{n} r_i S_{iA} & \sum_{i=1}^{n} r_i S_{iB} \end{bmatrix} \end{matrix}$$

Cue diagnosticity, $d_i = |S_{iA} - S_{iB}|$

Model B

Cue Diagnosticity

$$\begin{matrix} & \text{Hyp A} & \text{Hyp B} \\ & \begin{bmatrix} d_1 & \emptyset \\ \emptyset & d_2 \\ d_3 & \emptyset \\ & \vdots \\ \emptyset & d_n \end{bmatrix} & \\ \left[r_1, r_2, r_3, \ldots, r_n\right] & & = \begin{bmatrix} \sum_{i=1}^{n} r_i d_i & \sum_{i=1}^{n} r_i d_i \end{bmatrix} \end{matrix}$$

Figure 1: Decision Models: Model A represents a general adding-multiplying process; cue diagnosticity equals the absolute difference of the information value of each alternative hypothesis for a given cue; Model B represents a simplification in that only the cue diagnosticity is presented for the hypothesis having the greatest information value S_i.

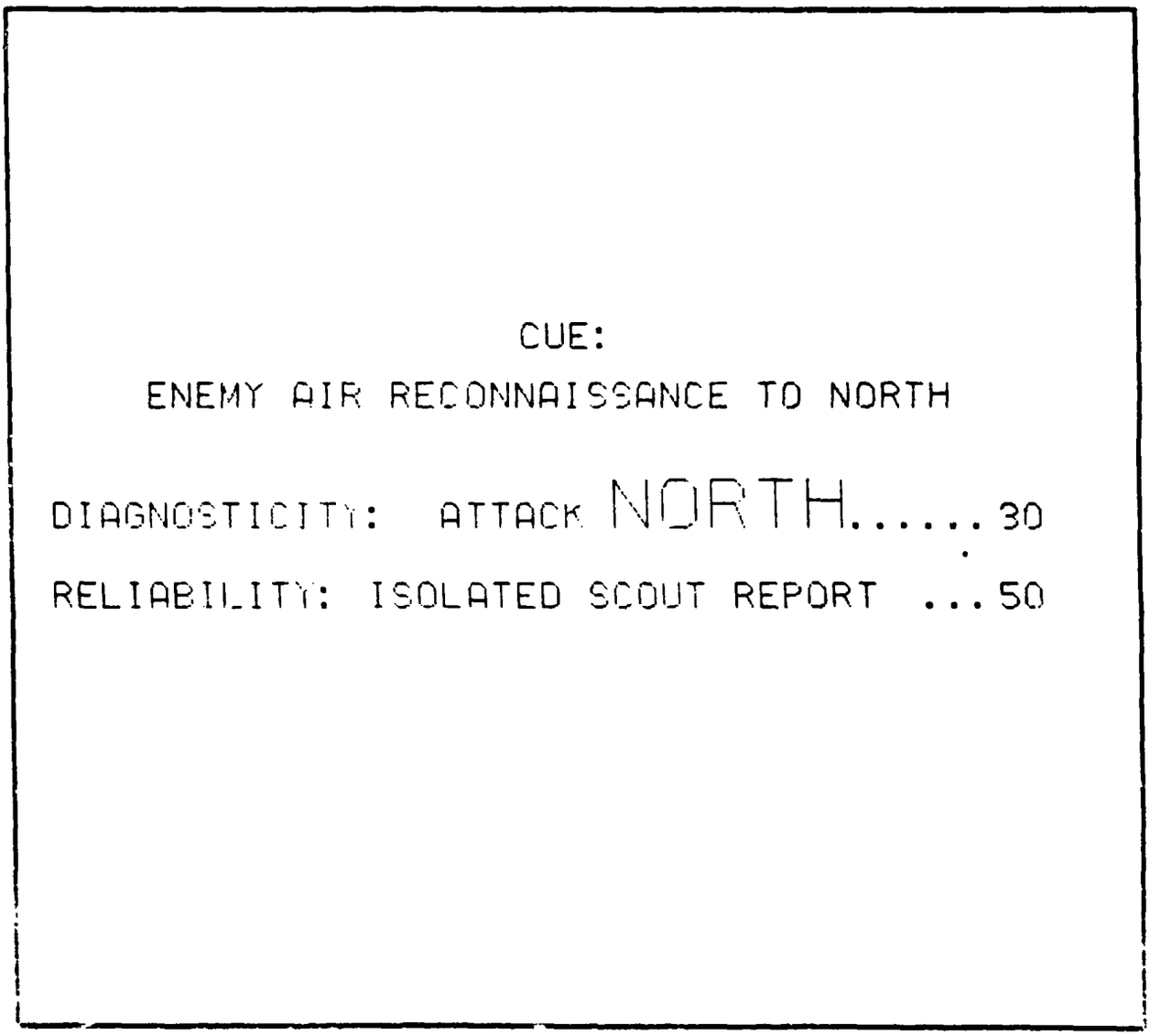


Figure 2: Information display with verbal code format.

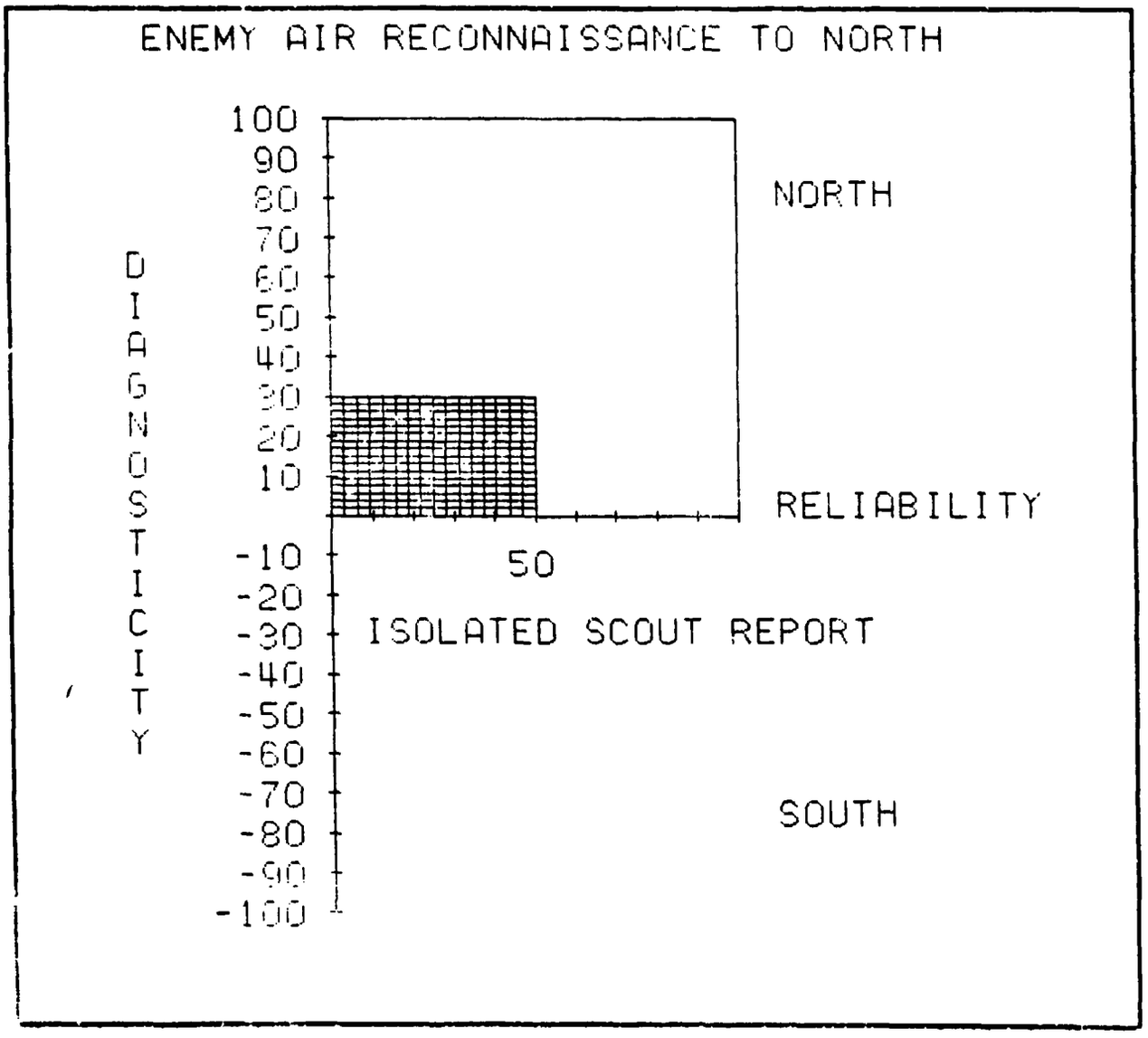


Figure 3: Information display with spatial code format.

3. The problem size or number of cues presented within a problem was varied. Two different effects were predicted. First, decision accuracy was expected to decrease with an increased problem size due to the effects of increased memory load. Secondly, confidence was predicted to increase with an increased problem size. This direct relationship is clearly supported in a vast number of studies but is particularly relevant to within subject designs (Kaplan & Major, 1973).

4. Cue variability, defined as the direction of correlation between cue diagnosticity and reliability, was varied in this experiment. Cues having a negative correlation between diagnosticity and reliability values, which we call high variability cues, were expected to be overestimated in the spatial format condition. As described above, we predict that the "perimeter effect" may produce a bias to overweight the information value of highly eccentric rectangles. If these findings bear on our study, then high variability cues will be overestimated in the spatial format condition. Correspondingly, if one of the two competing hypotheses is supported by high variability cues, the subjective value of support for this hypothesis will be overestimated. Because the perimeter effect is not relevant to the verbal code format condition, an overestimation bias is not expected in this condition. Therefore, a code x variability interaction is expected.

5. Finally, the total difference in information presented for the competing hypotheses within a trial was varied between problems. As stated above, the model of information integration presented in Figure 1 is incomplete. It describes the procedure for summing information for each hypothesis over successive cues, but does not describe how the final judgment on the difference of information is to be made. The preference and ratio models discussed above are relevant to this final judgment process. The preference model, where preference = $value_1 - value_2$, can clearly predict the most likely hypothesis, but intuitively does not describe a confidence judgment. A ratio model would describe confidence as a function of $V_1/(V_1 + V_2)$. Alternatively, the ratio model can describe the confidence rating but not the initial decision process of determining the most likely hypothesis. A combination of both models where weighted difference = $\frac{V_1}{V_1+V_2} - \frac{V_2}{V_1+V_2}$, or $\frac{V_1-V_2}{V_1+V_2}$ is evaluated. It was predicted that confidence will be directly related to the weighted difference factor. Optimality of information integration can be judged by the extent to which judged confidence covaries with the actual weighted (or absolute) difference.

Experiment 2

The objective of this experiment was to determine if there are biases associated with processing the diagnosticity and reliability stimulus dimensions. Hence, cues of high diagnosticity and low reliability were

consistently presented for one hypothesis and cues of low diagnosticity and high reliability presented for the competing hypothesis. If subjects tend to treat reliability as a categorical, overweighted variable then the hypothesis supported by evidence for which objective reliability is low should be systematically favored.

Method: Experiment 1

Subjects

Eight undergraduate students at the University of Illinois, four male and four female, volunteered to serve in this experiment. All subjects had normal or corrected vision and were paid $3.00 per hour for their participation in each of the three days of testing.

Apparatus

Subjects were seated in a light and sound attenuated booth containing a 10 cm x 8 cm Hewlett-Packard model 1330a CRT and two spring-return push-button keyboards. The keyboards at the right and left hands had two and ten buttons, respectively. Subjects sat approximately 90 cm from the display. A Digital Equipment Corporation PDP 11, 16 bit computer with 24K memory was used to generate the experimental displays and record subject performance.

Task

Figures 2 and 3 illustrate examples of the cues of military intelligence, a series of which were presented to subjects sequentially. Successive cues alternated in support of the two different possible courses of action that the threat force might take: attack North or attack South. Subjects were instructed to process the cues utilizing an adding-multiplying model. That is, cue valence was to be equal to diagnosticity x reliability and successive cue valences were to be summed in support of their respective hypotheses. Then at the completion of the trial subjects were prompted to assess the difference in support of the two alternative hypotheses and indicate which was the most likely enemy course of action using the two button keyboard. Subjects were then presented a confidence scale ranging from 0-9 and anchored by "absolutely uncertain" and "absolutely certain," respectively. Figure 4 illustrates this confidence scale. Subjects were instructed to assess how confident they were that the threat force would execute the course of action which the subjects had judged most likely given the available information. The ten button keyboard was used for this response.

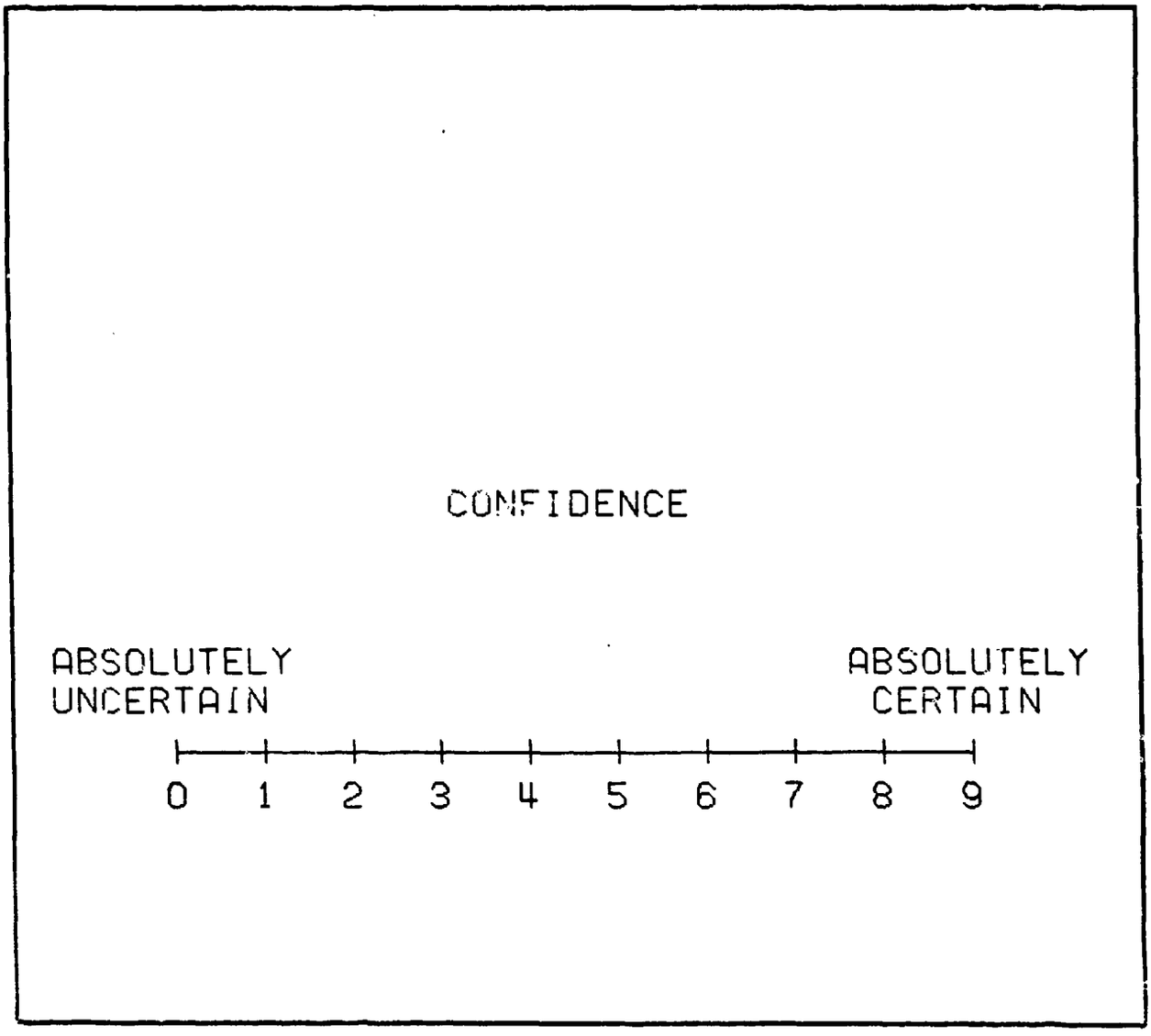


Figure 4: Confidence response scale.

Stimulus Design

Each cue consisted of a statement about current threat force disposition or relevant environmental conditions in the area of operations. Additionally, the source of the information was presented. Values of diagnosticity (the relevance of information) and reliability (the credibility of the source) were also presented as part of the cue. The following factors were manipulated orthogonally in the experimental design.

Coding. Two cue formats were evaluated. A verbal code fomat is presented in Figure 2. The alternative spatial code design utilizing dimensional integrality is presented in Figure 3. Cue valence is represented by the darkened rectangle where height is the diagnostic value and width is the reliability value. Position coding is utilized to indicate which of the two alternative hypotheses is supported by the cue. Information presented above the abscissa supports a threat attack to the North and information presented below supports a threat attack to the South.

Problem size. Three levels of problem size were used: A total of six cues, three in support of each hypothesis; eight cues total, four for each hypothesis; and 10 cues total, five for each hypothesis.

Trial variability (VARIABILITY). This factor concerns the relationship between cue diagnosticity and reliability dimensions. Problems containing only cues in which reliability and diagnosticity are positively correlated are designated low variability cues. Hence, most of the rectangles in the spatial format are "squarish." Cues with high diagnosticity and low reliability or low diagnosticity and high reliability are designated high variability cues. These cues have the form of eccentric rectangles in the spatial format. The matrix presented in Figure 5 depicts the respective diagnosticity and reliability values associated with high and low variability cues.

Trial variability has two levels. In the low case, both alternative hypotheses are supported with low variability cues. In the high trial variability case one hypothesis is supported with low variability cues and the second hypothesis is supported with high variability cues. Hence, if a perimeter bias is present, subjects will "overpredict" the second hypothesis. Note that cue variability is not a factor but a stimulus condition. Trial variability is the experimental factor having two levels and is dependent upon the cue variability condition presented to the two different alternative hypotheses during a trial.

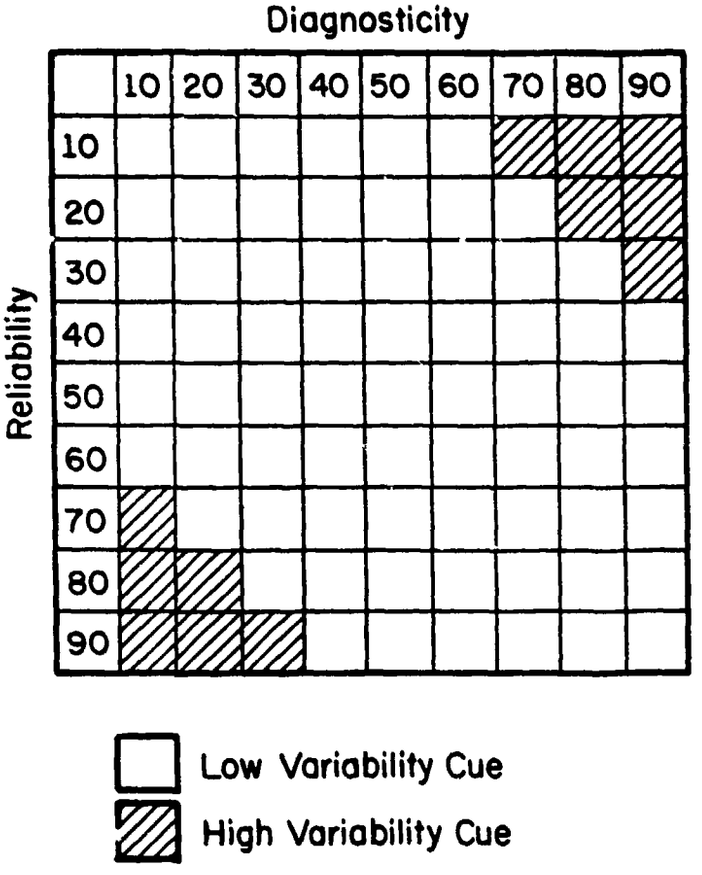


Figure 5: Cue variability design.

Presentation time (TIME). Two levels of cue presentation time were used. Cues were presented for three seconds (fast) or five seconds (slow). Interstimulus interval remained constant at one second.

Weighted difference. The weighted difference was computed by dividing the absolute difference of support presented for the two different hypotheses by the total support presented for both hypotheses. Hence, weighted difference =

$$\frac{\sum_{i=1}^{n} r_{iA}\, d_{iA} - \sum_{i=1}^{n} r_{iB}\, d_{iB}}{\sum_{i=1}^{n} r_{iA}\, d_{iA} + \sum_{i=1}^{n} r_{iB}\, d_{iB}} \times 100\%$$

Sets of problems having three equally spaced levels of weighted difference were used; 5-10%, 15-20%, and 25-30%.

Design

A within-subjects design was employed in which each subject participated in all experimental manipulations over a period of two sessions. Each session lasted approximately one hour and consisted of 36 trials. Figure 6 illustrates the total 72 trial conditions presented over two sessions.

Trial order within a session was blocked by variability and weighted difference (2 x 3). These six trials were presented randomly within a block. Additionally, the six blocks of each session were randomly ordered for each subject. The time factor was split by session such that all 36 trials of the first session were at the slow level and all 36 trials of the second session were at the fast level.

Procedure

Subjects were run individually. One practice session preceded the two experimental sessions and lasted 90 minutes. During the practice session subjects were presented the tactical scenario described earlier and in detail in Appendix A. Subjects were also given definitions of cue reliability and diagnosticity and examples of each as illustrated in Appendix B. Any military terminology used in the experiment was thoroughly explained. Finally, subjects were instructed on the use of the adding model for integrationg successive cues within a trial. A visual aid illustrated in Appendix C was used for instructing this process. During the remainder of the practice session, approximately 30 minutes,

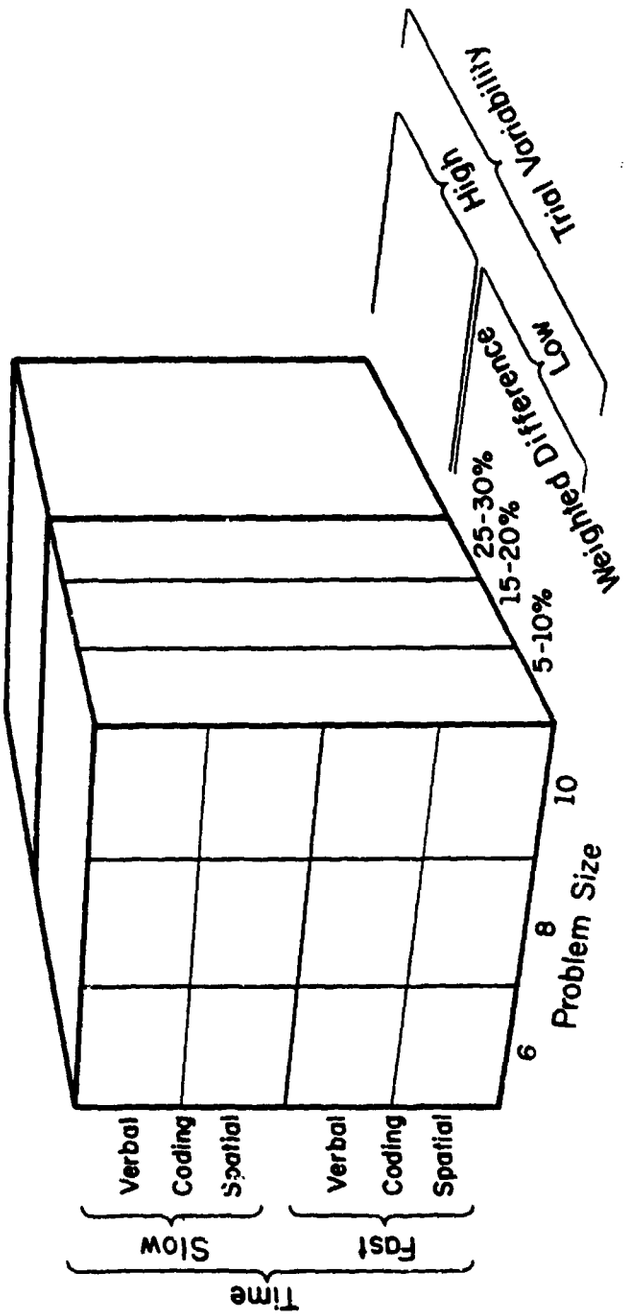


Figure 6: Experimental design.

practice trials were run. All subjects achieved a minimum of five successive correct trials during this period.

Subjects were given eight practice trials at the start of each of the two experimental sessions. A five second interval between trials and a 60 second rest between blocks of trials were imposed during the experimental sessions.

At the conclusion of the experiment subjects were asked to explain how they integrated successive cues in a trial and how their confidence rating was related to the information integration process.

Results: Experiment 1

Analysis of Experimental Effects

Accuracy. Table 1 presents a summary of decision accuracy as a function of each of the five independent variables averaged over replications. Arc sine transformations were performed on accuracy percentages. An analysis of variance performed on these data revealed significant main effects on decision accuracy for three of the five variables studied. No interaction effects were found to be statistically significant, and hence only main effects are shown in the table. A large effect was found for coding ($t(1,7) = 3.85$, $p < .025$).[1] As predicted, the spatial code format yielded an improvement in decision accuracy over the verbal format. The effect of trial variability on decision accuracy was also statistically significant ($t(1,7) = 2.17$, $p < .05$). Decision accuracy was best in the low variability condition, and poorer when one hypothesis was supported by the more eccentric rectangles. Finally, the main effect of time was statistically significant ($t(1,7) = 4.27$, $p < .005$). Decision accuracy was greater in the fast presentation condition (3 seconds) than in the slow presentation condition (5 seconds).

In interpreting the time effect, it should be recalled that the main effect of time was not of interest in this study. The interaction effects of time were of primary concern. For this reason the two levels of time were split between sessions. All slow conditions were run in the first session and all fast conditions were run in the second session. Consequently, the time effect is possibly the result of a practice artifact. It is also possible however that the time manipulation resulted in memory effect rather than a time stress effect. That is, at the slower 5 sec cue presentation time, the loss of information about each cue due to memory accounts for a large portion of the process limitations. It is therefore plausible that a lesser degree of memory decay can account for the increase in decision accuracy at the fast cue presentation time.

[1] T-tests rather than F-tests were used to examine the three two-level main effects.

Table 1

Percent Accuracy

Trial Variability	Low	High	
	97.2%	92.3%	
Coding	Spatial	Verbal	
	97.6%	92.0%	
Time	Slow	Fast	
	93.4%	96.2%	
Weighted Difference	5-10%	15-20%	25-30%
	94.2%	95.3%	94.8%
Set Size	6(total)	8(total)	10(total)
	95.3%	95.3%	93.8%

Confidence. Raw confidence data were transformed using the following algorithm: If the decision response is correct then absolute confidence = 10 + confidence rating; if the decision response is incorrect then absolute confidence = 10 - confidence rating. This algorithm adjusts the range of confidence from 1 to 19. This transformation is made to penalize subjects more heavily if they made an error (chose the incorrect hypothesis) when they were extremely confident of being correct. Since roughly 95% of the responses were correct, most of the transformed values are greater than 10. Table 2 presents the data summary for the transformed confidence scores.

A six-way repeated measures ANOVA (code x weighted difference x time x problem size x variability x replication) was performed on the transformed confidence data. The main effects of code and problem size were not significant ($F(1,7) = 2.47$, $p < .16$) and ($F(2,14) = 0.78$, $p = .48$), respectively. Significant effects were found for evidence, time and trial variability, and for the code x time and the problem size x time interactions.

A very large effect of weighted difference was found ($F(2,14) = 25.10$, $p < .0001$). This effect is particularly important for two reasons. First, it demonstrates that subjects are extracting more information when more information is available. In addition, it demonstrates that subjects are using the confidence response scale as an analog for evidence.

Figure 7 portrays the two main interactions on the confidence variable that were observed. The figure shows confidence with the verbal display on the left panel and the spatial on the right. The abscissa within each panel represents the effects of problem size, and the two functions are those for the slow(dashed line) and fast(solid line) speed, respectively. Each data point is collasped across the three levels of weighted difference.

Across both panels we find that the faster speed generated reliably higher confidence ($F1,7 = 5.88$, $p < .05$). While as noted in discussing the analogous effect on accuracy, this might reflect the influence of practice, it might also be related to the effect of memory decay. The faster rate produces a smaller loss of information during integration and therefore warrants higher confidence in the final response. This interpretation is supported to some degree by the reliable interaction between speed and problem size ($F(2,14) = 5.97$, $p < .01$). Examining Figure 7 suggests that the major source of this interaction is between the 8 and 10 cue problems. Increasing problem size from 8-10 increases confidence at the fast rate but diminishes it at the slow rate. This suggests that two factors may be operating with changes in response speed. When the rate is slow, a good deal of forgetting of earlier cues takes

Table 2

Absolute Confidence

Trial Variability	Low 14.95	High 13.37	
Coding	Verbal 13.92	Spatial 14.41	
Time	Slow(5 sec) 13.93	Fast(3 sec) 14.40	
Weighted Difference	Low(5-10%) 12.76	Med(15-20%) 14.14	High(25-30%) 15.60
Set Size	6(total) 14.0	8(total) 14.26	10(total) 14.23

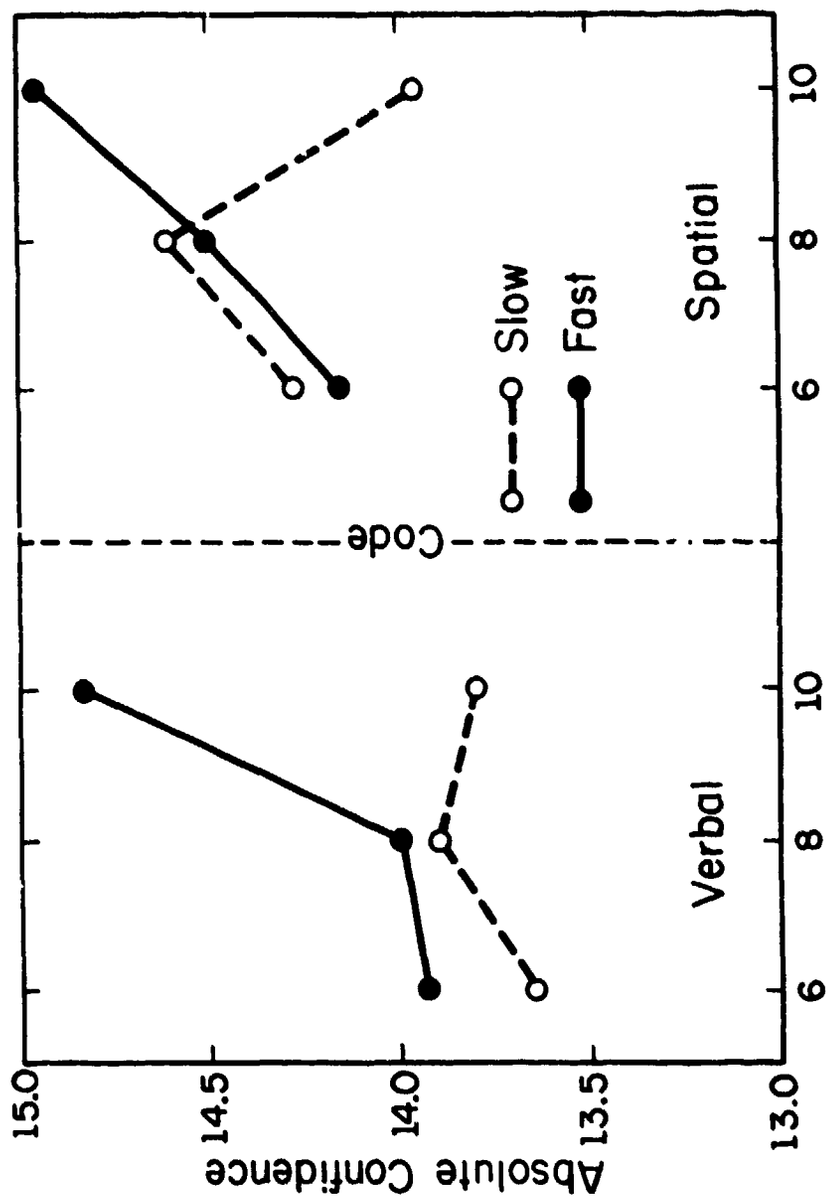


Figure 7: Code, speed, and problem size effects on confidence.

place and so, with more cues present more is forgotten and confidence is thereby reduced. This is substantiated by the marked loss in accuracy of this particular condition, there being twice as many errors here as in any of the other conditions. When the rate is fast there is less opportunity for decay. Here more cues lead to increased confidence because of the subjective belief that the evidence is more reliable.

An additional interaction effect of time x code was found to be statistically significant ($F(1,7) = 6.34$, $p < .04$) and is illustrated in Figure 7 as well. The verbal display shows a greater increase in confidence with faster speed than does the spatial which appears to be little affected by confidence at all, particularly with the small sized problems. There are two likely interpretations, the first of which corresponds to the practice artifact interpretation of the main effect of time. The information integration task is process limited and aided by the integral spatial display. The benefits of the spatial format over the verbal format decrease with practice (the faster speed) as the process limitations are decreased. Phrased in terms of our initial prediction, this interpretation suggests that the S-C-R incompatibility between the verbal cue format and spatial central processing code is somewhat overcome with practice. Alternatively, all or part of the process limitations may be memory related and the effects of the integral spatial display are largest when the greatest demands are placed on memory, which is likely to occur at the slow presentation rate.

A large main effect of trial variability was obtained ($F(1,7) = 22.47$, $p < .005$). Confidence ratings were higher in low variability trials than in high variability trials. This finding was predicted in the spatial code condition to result from an overestimation of high variability cues if perimeter estimates biased the computation of cue valence. In the present study, 24 of the 36 high variability trials were designed such that an overestimation of high variability cues would result in a confidence decrement. That is, the incorrect hypothesis was supported by high variability cues. The code x trial variability interaction was not statistically significant ($F(1,7) = .01$, $p = .94$). This indicates that if high variability cues are overestimated there is not a differential effect between the spatial code and verbal code format. Therefore, the overestimation bias appears to be operating in both the spatial and verbal code conditions, and so could not result from a "perimeter effect." Hence, it appears that high trial variability simply makes the information integration more difficult, with a resulting loss in both accuracy and confidence.

Regression analysis. The large effect of weighted difference on confidence discussed above demonstrates that subjects are treating the confidence scale as an analog of an evidence or information factor. We assume that the good decision maker is one who gains confidence as a greater difference in evidence between the competing hypotheses exists.

Additionally, the decision maker is "absolutely uncertain" when equal evidence is presented for both competing hypotheses. The following analysis investigates the relationship between evidence and confidence ratings, and the effects of our other independent variables on this relationship.

In order to capture a measure of how confidence varied with evidence, regression analyses of confidence on the three levels of weighted difference were performed for each subject in each of the 24 trial condition cells (code(2) x time(2) x problem size(3) x trial variability(2)). A five-way repeated measures ANOVA (code x time x problem size x trial variability x replication) was performed on the slope, intercept, and residual mean square statistics.

The main effect of coding was our primary interest in the ANOVA performed on slope. We have suggested that the spatial code format is more S-C compatible with an analog confidence scale. A difference in sensitivity or slope would therefore be of interest. This effect, however, was not statistically significant ($F(1,7) = .25$, $p = .63$), nor were any other main effects found to be significant.

The intercept statistic was interpreted as a measure of overconfidence. That is, we interpret a positive confidence estimate when the support for both hypotheses is extrapolated to be equal as a measure of overconfidence. Optimal decision makers are "absolutely uncertain" in this situation. This ANOVA showed a significant main effect of trial variability ($F(1,7) = 8.44$, $p < .02$). The mean intercept of the low variability trials (2.10) was greater than that of the high variability trials (0.78), a difference of 1.32. A main effect of trial variability on confidence was described earlier and again, Table 2 illustrates that the mean confidence rating of low trial variability was 1.58 units greater than the mean confidence rating of the high trial variability condition. We can now interpret this difference as related to a bias in assigning confidence between the two conditions--a bias that is somewhat unrelated to the actual differences in evidence.

The final ANOVA was performed on the residual mean square statistic. We interpret the residual mean square as a measure of "goodness of fit" of the regression line. Any conditions having significantly large residual mean square values would be an indication that performance was non-optimal or possibly that the weighted difference model was inappropriate as a descriptive model of confidence rating in this condition. The ANOVA, however, showed no significant effects.

It is important to note that the absence of reliable effects in the regression analyses may result in part from the lack of power associated with the three ANOVA's. Each individual regression plot of each cell, from which the statistics were derived, had only three data points. It is quite likely then that a single aberrant data point could drastically influence the regression slope producing a high degree of variability in the raw data. An alternative approach, in which data are averaged across subjects before computing the regression equation will be described below.

Order effects. If changing the order of stimulus presentation in a sequential integration task results in an altered response, then in general, order effects are present. There are many possible causes of order effects. Primacy effects are found when earlier stimuli in a serial integration task are given relatively more weight than later stimuli. This is a manifestation of the phenomenon of anchoring in which an initial hypothesis is accepted, based upon the first arriving evidence, and is then held more tenaciously than warranted when evidence arrives to disconfirm it (Kahneman & Tversky, 1973; Lopes, 1982). Correspondingly, recency effects are found when later stimuli are given more weight than earlier stimuli. The most common explanation of the recency effect is that the earlier stimuli are given less weight due to a memory loss. The recency effect seems particularly relevant to this study. This is because the role of memory loss was identified as a possible cause of of the significant interaction between time and problem size. This was attributed to the influence of memory loss in the 10 cue, T(slow) condition.

In the present study the cues systematically alternated in favor of one then the other hypothesis. The order of alternation was balanced so that on half of the trials ("correct first") the evidence concerning the correct hypothesis was presented as the first cue and that concerning the incorrect as the last cue. On the other half ("correct last"), evidence for the incorrect hypothesis was presented first and for the correct was on the last cue. Hence, if primacy were a dominant factor, then accuracy (and confidence) on correct-first trials should be greater than on correct-last trials. On the other hand, if information integration was dominated by recency, then correct-last trials should be favored. A two-way repeated measures ANOVA was performed on the confidence measures, and this effect was not found to be statistically significant ($F(1,7) = 1.24$, $p = .30$). The absence of an effect here does not necessarily mean that primacy and recency were not shown. It does imply that if such effects were present they probably balanced each other in their magnitude. The problems were not ordered in such a way as to choose between these particular hypotheses.

Model of Information Integration

Processing strategies (self-report). Subjects were asked to explain their information integration strategy for the entire decision task. More specifically, three processes were of interest: the assessment of the valence of each individual cue, the integration of successive cues, and the final assessment of confidence.

All subjects reported using a multiplying model, i.e., diagnosticity x reliability, to assess cue valence. The strategies for integration of successive cues reported by the subject were less unanimous. Three subjects reported using an adding model in which running totals were kept throughout a trial for each alternative hypothesis. Four subjects reported using a "random walk" model. Successive cues were added (or subtracted) to a single running balance. This model intuitively places less demands on working memory. Additionally, one subject reported using a somewhat primitive heuristic in which the "random walk" model was utilized, but was further simplified by using fingers and fractions of fingers to represent the running balance. Finally, subjects were asked to explain their strategy for assessing confidence. Five subjects stated that they used a weighted difference strategy. Three subjects claimed to use only the absolute difference of information presented for the two alternative hypotheses.

Regression analysis. The results of the self-report exercise have led us to take a closer look at the model subjects used for information integration. Three subjects reported using a preference model where net differences = $value_1$ - $value_2$, in favor of the weighted difference model. These two models are qualitatively very different. Recall that in this experiment successive cues alternate in support of the two alternative hypotheses. An equal number of cues are presented for each hypothesis. The weighted difference model does not have directional constraints. If one hypothesis is strongly supported and two additional cues of equal weight are presented for each alternative hypothesis, so that the net evidence of these two cues is zero, then a confidence decrement is predicted. This makes the weighted difference model qualitatively similar to an averaging model. Alternatively, the net difference model has directional constraints akin to the Bayesian model. This model does not predict a decrease in response when additional neutral or mild evidence is presented.

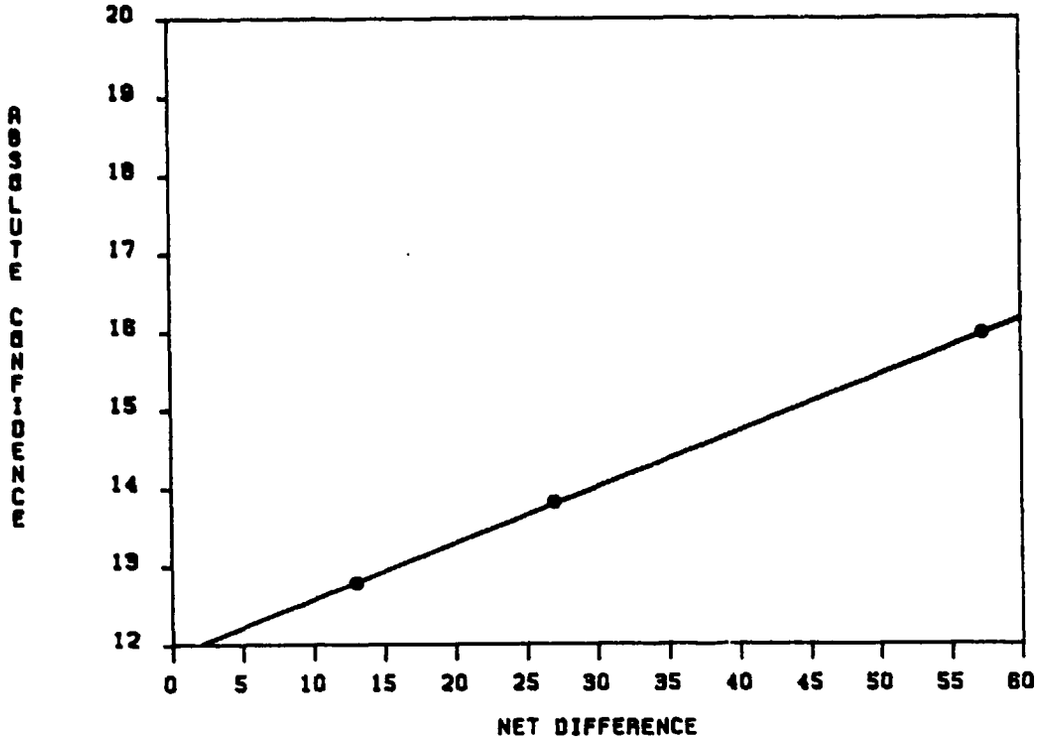
Since we did not "track" the confidence judgment of our subjects as they progressed through the sequence of cues (only a single rating was given after all cues were presented), we were unable to distinguish between the models on the basis of the impact of weak or neutral evidence. Instead, we employed a means of model testing that capitalized on the fact that we had three different problem sizes, defining three different levels

of total evidence. As described in the methods section, the three levels of evidence of our cues were designed to create equally spaced equal intervals of weighted difference (evidence for the favored hypotheses/total evidence). If subjects then used this variable to calibrate their subjective confidence, then confidence should vary linearly with net difference. On the other hand, the three objective evidence levels were not equally spaced in terms of the net difference (equal values of relaive differences will produce different values of net difference if the total amount of evidence varies). Hence, if subjects used the more optimal net difference strategy, the regression of subjective confidence onto relative confidence should show a non-linear component.

Regressions of confidence on net difference, and confidence on percent difference were performed for each subject in 24 different trial conditions (code(2) x time(2) x problem size(3) x trial variability(2)). In order to determine if subjects were using one mode of information integration over another, the objective evidence for the favored hypothesis on each trial was computed in two different fashions. First, as the net evidence in favor of one over the other; second, as the ratio of this net difference to the total amount of evidence presented for both hypotheses on the trial. A two-way (decision model x replication) repeated measures ANOVA was performed on the correlation statistics of the two different regressions. Correlations were interpreted as a degree of relationship measure between confidence ratings and the respective model of integration. No significant effect of correlation for the two different models was found ($F(1,7) = .77$, $p = .41$).

One final analysis was conducted to evaluate the difference between these two models. Figure 8 illustrates the regression of the mean confidence rating on net difference, and below that the regression of confidence on percent or relative difference. The mean percent difference was computed in each of the three levels of weighted difference and plotted against the respective mean confidence ratings over subjects. A similar procedure was used to plot confidence on net difference. Note that the ordinate (confidence) values are equivalent between the two graphs (despite the expanded scale on the bottom), but the abscissa values are equally spaced on the weighted difference scale (as we had created them), while they are not on the net difference scale.

As Figure 8 indicates, both models seem to do an excellent job of predicting confidence with averaged group data. There is however, an apparently better relationship depicted in the plot of confidence on net difference. This is of course, only a cursory analysis. Recall, however, that weighted difference and net difference are highly related. This small difference may therefore yield substantial evidence in favor of the more optimal net difference model.



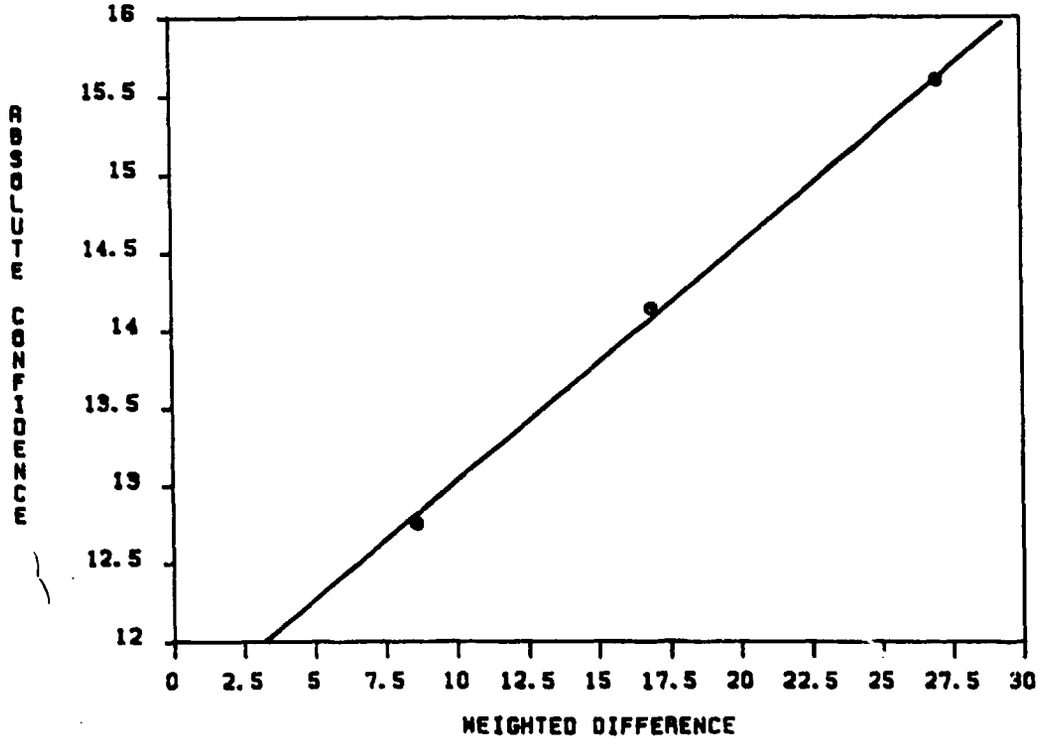


Figure 8: Comparison of net difference and weighted difference as predictors of confidence.

The data were further broken down and plotted in the same fashion separately for the two groups that differed in the strategies that they had reported using. These self-reports proved to be fairly accurate indicators of the integration strategy revealed in the data. The three subjects who reported that they used the net difference strategy showed an almost perfect linear relation between net difference and subjective confidence. The fit was much poorer with the weighted difference. The behavior of the five self-reported weighted difference subjects on the other hand seemed to reflect a compromise between the two strategies.

Method: Experiment 2

The design of Experiment 2 was similar to that of Experiment 1 with three exceptions. First, the problem size remained constant at eight total cues, four for each alternative hypothesis. Second, the display time was held constant at the slow (5 seconds) presentation rate. Finally, a new level of trial variability was implemented. All trials in this experiment utilized high variability cues for both alternative hypotheses. More specifically, cues of high diagnosticity and low reliability were presented for one hypothesis and cues of low diagnosticity and high reliability were presented for the alternative hypothesis. Formally, in terms of any model of information integration neither hypothesis should be favored by this bias since the valence of each cue should be insensitive to the relative contributions of reliability and diagnosticity (Johnson, Cavanagh, Spooner, & Samet, 1973). However, if subjects tended to treat one variable different from another (i.e., to discount differences in reliability applying the "as if" heuristic; Wickens, 1983), then biases should become evident. The trials of Experiment 2 were configured so that the incorrect hypothesis was always favored by the cues of low reliability. Hence, to the extent that subjects overestimated reliability these inflated cue valences should bias them toward picking the incorrect hypothesis--i.e., their error rate should increase.

Results: Experiment 2

Accuracy. Table 3 presents accuracy data averaged over subjects and over display format. The accurrcy measure for each subject was determined by the proportion of errors made with the two opportunities (one with each display type) in a given condition. Hence, raw accuracy values for a given subject were either 0, .5, or 1. Arc sine transformations were performed on accuracy percentages. A two-way (weighted difference x replication) repeated measures ANOVA was conducted. As in Experiment 1, the main effect of weighted difference was statistically significant ($F(1,7) = 5.86$, $p < .01$). It is evident that decision accuracy was poorest in the trials of low weighted difference. As described above, the evidence in all trials favored the hypothesis having low diagnosticity and high reliability. Therefore, the fact that in 44% of all low difference trials the subject chose the incorrect hypothesis, supported by high

diagnosticity and low reliability, indicates that reliability tends to be overweighted to a greater extent than diagnosticity. The 44% error rate in this condition is significant in that it compares with an error rate of less than 6% in the corresponding condition of Experiment 1, when conditions were not created to induce a bias. While the present bias might be expected to be present at all three levels of weighted difference, its presence here only in the low weighted difference condition is not too surprising because a constant bias would have a relatively smaller effect on the greater weighted difference.

Confidence. Because we have demonstrated that the confidence scale is used as an analog of evidence we expect the dimensional bias for reliability to decrease confidence at all three levels of weighted difference. That is, since the cues in support of the incorrect hypothesis (high diagnosticity and low reliability) are overestimated, this would yield a decrease in confidence as predicted by the weighted difference model.

dgments of confidence were transformed as in Experiment 1. Table 3 presents these data averaged across subjects. A three-way (code x weighted difference x replication) ANOVA was conducted. The main effect of code and the code x weighted difference interaction were not significant, supporting the interpretations made for the absence of these effects in the accuracy ANOVA. The main effect on weighted difference was very large ($F(7,14) = 18.84$, $p < .0001$). This demonstrates that, as in Experiment , subjects are becoming increasingly confident as a greater difference in evidence between the competing hypotheses exists. Most significantly for the hypothesis under consideration, the mean confidence rating i his experiment was 12.69 on the experimental confidence scale. The ove confidence for the corresponding eight cue conditions in Experime 1 was 14.16. The decrease in confidence in Experiment 2 supports prediction of a dimensional overestimation bias, i.e., cues of high diagnosticity and low reliability are overestimated, pulling the integration of information away from the correct hypothesis and towards the neutral int, and hence producing a final judgment of reduced confidence.

Discussion

Engineering Applications

Code effects. The results of Experiment 1 indicate that decision accuracy is enhanced with the integrated spatial display format. Two interpretations of this finding have been discussed. We initially predicted that the integral dimensions of the spatial code format display would simplify the integration process and, therefore, enhance decision accuracy, especially under conditions of time stress. This latter interpretation is now uncertain. The code x time interaction effect demonstrates that the spatial code format has a greater benefit when information is presented at a relatively slower rate. Alternatively, the

Table 3

Percent Accuracy

		Weighted Difference		
		5-10%	15-20%	25-30%
		0.563	0.938	1.000
		Absolute Confidence		
		Weighted Difference		
		5-10%	15-20%	25-30%
Code	Verbal	11.00	12.25	14.38
	Spatial	10.38	13.50	14.62

spatial code format was predicted to be more compatible with the analog nature of the internal scale of confidence maintained in working memory. Because of the greater degree of compatibility, the internal scale is more easily revised and is less sensitive to the decay of positioning with the increasing interval between successive cues. The spatial display was predicted to enhance the encoding process and have positive effects on working memory. The code x time interaction effects support this interpretation.

Speed effects. It was our intention to evaluate the effect of the spatial code format under conditions of time stress. Our results indicate, however, that the time manipulations in this study produced more of an effect on memory loss effect than of time stress. Both the code x time and problem size x time interaction have been interpreted in terms of memory loss, although these could also result from practice effects because the slow speed was presented first.

Non-optimality in Decision Making

In addition to their human engineering implications, the results of the present experiment bear as well on certain cognitive phenomena in information integration and decision making. Three of these phenomena will be described.

1. Models of information integration. The consistent effects of weighted difference on confidence in both experiments demonstrate the ability of the subjects to extract more evidence and therefore increase their confidence as more diagnostic evidence is presented. Five subjects did in fact report using a weighted difference model for confidence judgment. On the other hand, three subjects claimed to consider the net difference of information presented for the two competing hypotheses. It is unclear as to which of these two models, weighted difference or net difference, is the best predictor of subjective confidence. An ANOVA performed on the correlations of both models and confidence was inconclusive. A comparison of the regression plots of confidence on both of these models does, however, suggest that the net difference is the better overall predictor of confidence for group data. The plot for the three subjects who claimed to use the net difference was perfectly linear with this variable. The plot for the remaining five subjects could be equally well accounted for by either of the two models. Since the net difference model is more appropriate in the Bayesian type inference task used here than is the weighted model (reflecting an averaging process), we conclude that subjects in the present paradigm showed tendencies toward this form of optimal behavior. As noted of course, we were unable to assess the different models directly in terms of the impact of different cues presented in sequence. Future research will address this issue.

2. Individual cue values. Examining the integration of the individual dimensions of reliability and diagnosticity in a finer grain revealed two further effects. Experiment 1 demonstrated that a negative correlation between these variables (producing for the spatial display an increase in shape variability) reduced both the accuracy and confidence of prediction. The fact that this reduction was a main effect that did not differ between the verbal and spatial formats suggests that the source of difficulty was not the physical variability in the shape of the rectangles, but rather was related to problems encountered in integrating negatively correlated data.

Individual cue values were also examined in Experiment 2 whose data suggested that subjects tended to over-value low levels of reliability, thereby reducing both their accuracy and confidence, relative to the values observed in Experiment 1. These results are important in that they indicate that subjects truly did respond to the meaning of the two cues and did not simply treat them as arbitrary numbers. Had arbitrary numbers been combined there would be little reason for subjects to treat one differently from the other. In the spatial condition it is of course possible that the asymmetry was related to the physical dimensions. The same data would have been produced had subjects overestimated the base of the rectangles (depicting reliability) relative to the height. Yet two factors indicate that this did not occur. On the one hand, this bias would be contrary to the bias typically observed in the horizontal-vertical illusion in which vertical segments are overestimated in length relative to horizontal ones. On the other hand, since the main effect of code was not significant in Experiment 2, this would suggest an equal loss of judgment for both verbal and spatial displays, indicating that the source was related to cognitive information integration and not to perceptual display biases. Thus it appears that the subjects were processing the reliability measure as if its value were closer to that of the larger diagnosticity value. The general finding that differences in reliability are ignored or "unitized" in multiple source information integration tasks has been reported in a number of other investigations (see Wickens, 1983 for a summary). The demonstration of the "as if" heuristic here is consistent with those findings.

3. Serial effects. The problem size or number of cues within a trial did not have an effect on confidence. This is contrary to many findings of similar studies with serial integration. The absence of a problem size effect is indicative of optimal performance. If subjects have evaluated information to produce a tentative confidence rating, and then are presented additional information, they should not change their confidence rating unless the weighted difference of information changes.

Another indicator of optimal performance was the absence of order effects. Recall that, in accordance with our experimental design, one half of the trials presented the first cue of the trial in favor of the most likely hypothesis, while the last cue was in favor of the least likely hypothesis. Conversely, on the other half of the trial conditions, the last cue of the trial was presented in favor of the most likely hypothesis, while the first cue was not. Thus, a greater mean confidence in the former or latter trial conditions would indicate a primacy or recency effect, respectively. We did not find a significant effect although this must be interpreted with caution. It is certainly possible that neither effect is present, and hence the ANOVA did not show a significant effect. On the other hand, both primacy and recency may have been operating and simply cancelled each other out. In future work, an experimental design which does not employ a strictly alternating sequence of cue presentation will be used to clarify this somewhat ambiguous interpretation.

Implications and Future Research

Spatial code format displays should be considered for further study and application in tactical C^3 systems as well as other areas of decision making in which responses and the internal representations underlying those responses are analog in form. This display format seems particularly beneficial when decision performance is limited by memory loss. Additional research will be necessary to gain a better understanding of the effects of time stress on the utility of the spatial code format. Decreasing both the problem size and time of cue presentation would possibly limit the effects of memory loss and better define the effects of time stress on integration task performance.

The consistent effects of weighted difference on confidence ratings demonstrates the efficacy of this model for describing decision performance in the present paradigm. A simple net difference model however does just as well, if not better, than the weighted difference model in describing the amount of subjective evidence extracted by a subject in a decision trial. Further research will be necessary to clarify the relevance of these two descriptive models. It is hoped that an experimental design which varies trial evidence in accordance with both models, but in an orthogonal manner, would isolate the effects of both models.

Finally, additional insight on how the "as if" heuristic operates on different sources of information is required. We have good reason to believe that the operation of this heuristic on cues with high diagnosticity and low reliability results in an overestimation bias of the cue valence or worth. This "riskiness" in interpreting unreliable data is possibly the result of a cognitive simplification. The operator reduces the load imposed on working memory by ignoring reliability or placing its value at unity. An experimental design in which these cues, high in diagnosticity and low in reliabilty, are pitted against cues of equal total evidence, but of moderate diagnosticity and reliability, might demonstrate the exact nature of this bias. Further knowledge of this effect could have a great impact on C^3 system performance if considered in both training and display design.

References

Anderson, N. H. & Weis, D. J. Test of a multiplying model for estimated area of rectangles. American Journal of Psychology, 1971, 84, 543-548.

Bar Hillel, M. The base-rate fallacy in probability judgments. Acta Psychologica, 1980, 44, 211-233.

Dawes, R. M. The robust beauty of improper linear models in decision making. American Psychologist, 1979, 34, 571-582.

Dawes, R. M. & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Einhorn, H. J. & Hogarth, R. M. Behavioral decision theory. Annual Review of Psychology, 1981, 32, 53-88.

Fleming, R. A. The processing of conflicting information in a simulated tactical decision-making task. Human Factors, 1970, 12, 375-385.

Garner, W. R. The Processing of Information and Structure. New York: John Wiley & Sons, 1974.

Garner, W. R. & Felfoldy, G. L. Integrality of stimulus dimensions in various types of information processing. Cognitive Psychology, 1971, 1, 225-241.

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. Psychological Bulletin, 1970, 73, 422-432.

Johnson, E. M., Cavanagh, C., Spooner, R. L., & Samet, M. G. Utilization of reliability measurements in Bayesian inference: models and human performance. IEEE Transactions on Reliability, 1973, 22, 176-183.

Kahneman, D. & Tversky, A. On the psychology of prediction. Psychological Review, 1973, 80, 237-351.

Kanarick, A., Huntington, J., & Peterson, R. C. Multi-source information acquisition with optimal stopping. Human Factors, 1969, 71, 379-386.

Kaplan, M. F. & Major, G. Will you like me at set-size 3 as you might at 6?: Amount of information and attraction. Paper presented at the Meeting of the Psychonomic Society, St. Louis, MO, November, 1973.

Lockhead, G. R. Holistic vs. analytic process models: A reply. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 746-754.

Loo, R. Individual difference dimensions as human factors considerations in tactical communications systems. Proceedings, 14th Annual Meeting of the Toronto Human Factors Association of Canada, Tornoto, 1981.

Lopes, L. L. Toward a procedural theory of judgment. Technical Report, Wisconsin Human Information Processing Program (WHIPP 17), Madison, WI, December, 1982.

Lyon, D. & Slovic, P. Dominance of accuracy information and neglect of base rates in probability estimation. Acta Psychologica, 1976, 40, 287-289.

Samet, M. G. et al. Application of adaptive models to information selection in C systems. Perceptronics Technical Report PTR-1033-76-12, December, 1976.

Schum, D. The weighting of testimony of judicial proceedings from sources having reduced credibility. Human Factors, 1975, 17, 172-203.

Slovic, P., Fischhoff, B., & Lichtenstein, S. Behavioral decision theory. Annual Review of Psychology, 1977, 28, 1-39.

Smith, J. P. The effects of figural shape on the perception of area. Perception and Psychophysics, 1969, 5, 49-52.

Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 85, 1124-1131.

Wallsten, T. S. Processing information for decisions. In N. J. Castellan, D. B. Pisoni, & G. Potts (Eds.), Cognitive Theory (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum, 1977.

Wallsten, T. S. Processes and models to describe choice and inference. In T. S. Wallsten (Ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, NJ: Lawrence Erlbaum, 1980.

Wickens, C. D. Engineering Psychology and Human Performance. Charles Merrill, Columbus, Ohio, 1983.

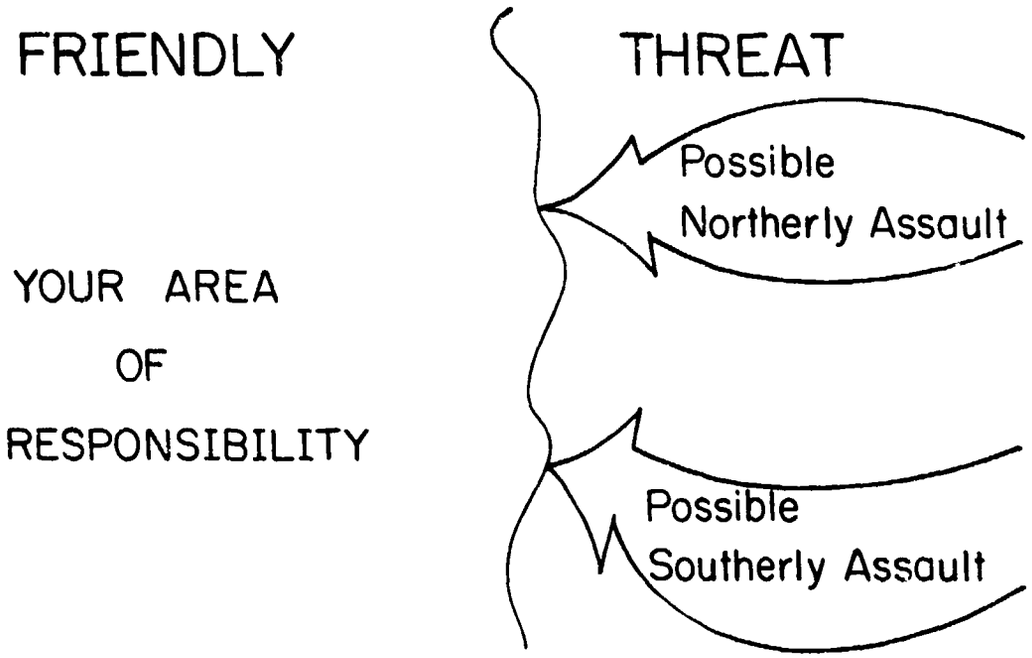
Wickens, C. D., Sandry, D., & Vidulich, M. Compatibility and resource competition between modalities of input, output, and central processing: Testing a model of complex task performance. Human Factors (in press).

Appendix A

Visual Model of Information Integration

A Simulated Tactical Decision Making Task

Imagine that you are the commander of an Army unit. You have just been advised that the threat force is preparing an assault against your sector (area) of responsibiiity. Your sector is very large and you cannot possibly cover (defend) the entire area. You can, however, successively block the threat assault if you know where in your area he will attack and you subsequently concentrate your forces there, i.e., will the enemy attack from the North or from the South?



You will be presented several pieces of evidence or cues. Your task will be to weigh the cues presented and decide which alternative is most likely, i.e., will the threat assault in the North or South of the sector?

Appendix B

How Much Weight to Assign Each Cue

A cue is a piece of evidence which supports one of the possible alternatives. The weight or valence of a cue is a function of the cue diagnosticity and the cue reliability.

Cue diagnosticity is determined by the relevance of the information on the decision at hand. Example: Consider a jury weighting pieces of evidence (cues) in a murder trial.

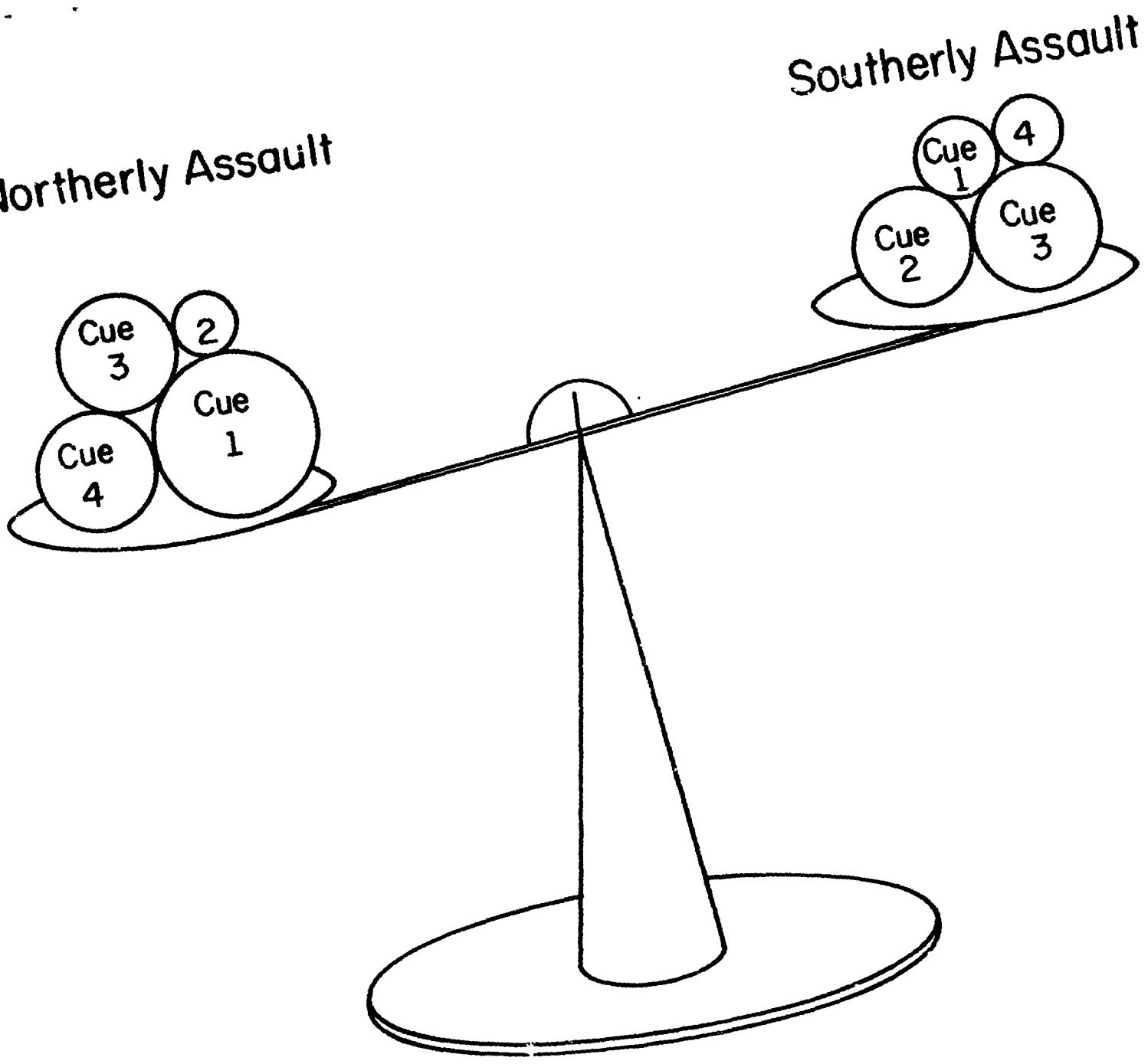
> Cue 1: A character witness has testified that the defendant has a "bad temper." This information is at best circumstantial and not very relevant. This cue would have a very low diagnostic weight, possibly 10 on a scale of 0-100.
>
> Cue 2 : The subject was seen fleeing the scene of the murder. This cue is very relevant and would surely implicate the defendant in the murder. This cue would have a very high diagnostic weight, possibly 90 on a scale of 0-100.

Cue reliability is determined by the credibility of the source of information. A source can be a person, thing, or activity from which the information was originally obtained. Example: Consider two different cases of Cue 2 above.

> Case 1: This defendant was seen fleeing the scene of the murder. The witness was a policeman responding to the call for help. The policeman would have great credibility and the reliability of this cue would be very high, possibly 90 on a scale of 0-100.
>
> Case 2: The subject was seen fleeing the scene of the murder. The witness is a known felon who is also suspect in the murder. This witness would have very little credibility. The reliability of this cue would be low, possibly 10 on a scale of 0-100.

Therefore, cue weight = diagnosticity x reliability. What weight would you assess the video replay of John Hinkley's assault on President Reagan if used as evidence in a trial?

Appendix C
Visual Model of Information Integration



OFFICE OF NAVAL RESEARCH

Engineering Psychology Group

TECHNICAL REPORTS DISTRIBUTION LIST

CAPT Paul R. Chatelier
Office of the Deputy Under
Secretary of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Engineering Psychology Group
Office of Naval Research
Code 442 EP
Arlington, VA 22217 (2 cys)

Aviation & Aerospace Technology
Programs
Code 210
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Physiology & Neuro Biology
Programs
Code 441NB
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Manpower, Personnel &
Training Programs
Code 270
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Mathematics Group
Code 411-MA
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Information Sciences Division
Code 433
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

CDR K. Hull
Code 230B
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
Arlington, VA 22217

Dr. J. Lester
ONR Detachment
495 Summer St.
Boston, MA 02210

Mr. R. Lawson
ONR Detachment
1030 E. Green St.
Pasadena, CA 91106

CDR James Offutt
Officer-In-Charge
ONR Detachment
1030 E. Green St.
Pasadena, CA 91106

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D.C. 20375

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. Robert Blanchard
Navy Personnel Research
and Development Ctr.
Command and Support Systems
San Diego, CA 92152

CDR J. Funaro
Human Factors Engr. Div.
Naval Air Development Ctr.
Warminster, PA 18974

Mr. Stephen Merriman
Human Factors Engr. Div.
Naval Air Development Ctr.
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons
China Lake, CA 93555

Human Factors Engr. Branch
Code 1226
Pacific Missile Test Ctr.
Point Mugu, CA 93042

Dr. S. Schiflett
Human Factors Section
Systems Engr. Test Directorate
U.S. Naval Air Test Center
Patuxent River, MD 20670

CDR C. Hutchins
Code 55
Naval Postgraduate School
Monterey, CA 93940

Office of the Chief of Naval
Operations (OP-115)
Washington, D.C. 20350

Dr. Edgar M. Johnson
Technical Director
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Director, Organizations &
Systems Research Lab
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Technical Director
U.S. Army Human Engr. Labs
Aberdeen Proving Ground, MD 21005

U.S. Air Force Office of
Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Chief, Systems Engr. Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks AFB, TX 78235

Dr. Daniel Kahneman
University of British Columbia
Department of Psychology
Vancouver, BC V6T 1W5
Canada

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsview, Ontario M3M 3B9
Canada

Dr. A.D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
ENGLAND

Defense Technical Information Ctr.
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 cys)

Dr. Craig Fields
Director, System Sciences Office
Defense Advanced Research Projects
AGency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. M. Montemerlo
Human Factors & Simulation
Technology, RTE-6
NASA HQS
Washington, D.C. 20546

Dr. Robert R. Mackie
Human Factors Research Division
Canyon Research Group
5775 Dawson Ave.
Goleta, CA 93017

Dr. Amos Tversky
Dept. of Psychology
Stanford University
Stanford, CA 94305

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Dr. T.B. Sheridan
Dept. of Mechanical Engr.
Massachusetts Institute of
Technology
Cambridge, MA 02139

Dr. Harry Snyder
Dept. of Industrial
Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, VA 24061

Dr. Robert T. Hennessy
NAS - National Research Council (COHF)
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Ave.
Woodland Hills, CA 91364

Dr. Robert C. Williges
Dept. of Ind. Engr. & OR
Virginia Polytechnic Institute
and State College
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Deborah Boehm-Davis
General Electric Company
Information Systems Programs
1755 Jefferson Davis Highway
Arlington, VA 22202

Dr. Charles Gettys
Dept. of Psychology
University of Oklahoma
455 W. Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Boulder, CO 80309

Dr. James H. Howard, Jr.
Dept. of Psychology
Catholic University
Washington, D.C. 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Dr. Edward R. Jones
Chief, Human Factors Engineering
McDonnell-Douglas Astronautics Co.
St. Louis Division
Box 516
St. Louis, MO 63166

Dr. Babur M. Pulat
Dept. of Industrial Engineering
North Carolina A&T State University
Greensboro, NC 27411

Dr. Lola Lopes
Information Sciences Division
Dept. of Psychology
University of Wisconsin
Madison, WI 53706

Dr. Stanley N. Roscoe
New Mexico State University
Box 5095
Las Cruces, NM 88003

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Executive Park
Burlington, MA 01803

Dr. William B. Rouse
School of Industrial & Systems Engr.
Georgia Institute of Technology
Atlanta, GA 30332

Dr. Richard Pew
Bolt Beranek & Newman, Inc.
50 Moulton St.
Cambridge, MA 02238

Dr. Hillel Einhorn
Graduate School of Business
University of Chicago
1101 E. 58th St.
Chicago, IL 60637

Dr. Douglas Towne
University of Southern California
Behavioral Technology Laboratory
3716 S. Hope St.
Los Angeles, CA 90007

Dr. Robert G. Smith
Office of the Chief
of Naval Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

Human Factors Department
Code N-71
Naval Training Equip. Ctr.
Orlando, FL 32813

CDR Norman E. Lane
Code N-7A
Naval Training Equip. Ctr.
Orlando, FL 32813

Dr. Gary Poock
Operations Research Dept.
Naval Postgraduate School
Monterey, CA 93940

Dr. Ross Pepper
Naval Ocean Systems Center
Hawaii Laboratory
P.O. Box 997
Kailua, HI 96734

Dr. A.L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

Dr. L. Chmura
Naval Research Laboratory
Code 7592
Computer Sciences & Systems
Washington, D.C. 20375

Human Factors Technology
Administrator
Office of Naval Technology
Code MAT 0722
800 N. Quincy St.
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 334A
Washington, D.C. 20361

Commander
Naval Air Systems Command
Crew Station Design
NAVAIR 5313
Washington, D.C. 20361

Mr. Philip Andrews
Naval Sea Systems Command
NAVSEA 03416
Washington, D.C. 20362

Commander
Naval Electronics Systems Command
Human Factors Engineering Branch
Code 81323
Washington, D.C. 20360

Larry Olmstead
Naval Surface Weapons Ctr.
NSWC/DL
Code N-32
Dahlgren, VA 22448

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Ctr.
Bethesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Dept.
Naval Medical Research Institute
Bethesda, MD 20014

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Commander, Naval Air Force
U.S. Pacific Fleet
Attn: Dr. James McGrath
Naval Air Station, North Island
San Diego, CA 92135

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Dr. Marshall Farr
Office of Naval Research
Code 442
800 N. Quincy St.
Arlington, VA 22217

Dr. Earl Hunt
Department of Psychology
University of Washington
Seattle, WA 98105